Lasso-type recovery of sparse representations for high-dimensional data

Nicolai Meinshausen and Bin Yu[*]

Department of Statistics, UC Berkeley

December 5, 2006

Abstract

The Lasso (Tibshirani, 1996) is an attractive technique for regularization and variable selection for high-dimensional data, where the number of predictor variables p is potentially much larger than the number of samples n. However, it was recently discovered (Zhao and Yu, 2006; Zou, 2005; Meinshausen and Bühlmann, 2006) that the sparsity pattern of the Lasso estimator can only be asymptotically identical to the true sparsity pattern if the design matrix satisfies the so-called *irrepresentable condition*. The latter condition can easily be violated in applications due to the presence of highly correlated variables.

Here we examine the behavior of the Lasso estimators if the *irrepresentable condition* is relaxed. Even though the Lasso cannot recover the correct sparsity pattern, we show that the estimator is still consistent in the ℓ_2-norm sense for fixed designs under conditions on (a) the number s_n of non-zero components of the vector β_n and (b) the minimal singular values of the design matrices that are induced by selecting of order s_n variables. The results are extended to vectors β in weak ℓ_q-balls with $0 < q < 1$. Our results imply that, with high probability, all important variables are selected. The set of selected variables is a useful (meaningful) reduction on the original set of variables ($p_n > n$). Finally, our results are illustrated with the detection of closely adjacent frequencies, a problem encountered in astrophysics.

Report Documentation Page		Form Approved OMB No. 0704-0188

1. REPORT DATE **05 DEC 2006**	2. REPORT TYPE	3. DATES COVERED **00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE **Lasso-type recovery of sparse representations for high-dimensional data**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Department of Statistics,Berkeley,CA,94720-3860**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The Lasso (Tibshirani, 1996) is an attractive technique for regularization and variable selection for high-dimensional data, where the number of predictor variables p is potentially much larger than the number of samples n. However, it was recently discovered (Zhao and Yu, 2006; Zou, 2005; Meinshausen and B?uhlmann, 2006) that the sparsity pattern of the Lasso estimator can only be asymptotically identical to the true sparsity pattern if the design matrix satisfies the so-called irrepresentable condition. The latter condition can easily be violated in applications due to the presence of highly correlated variables. Here we examine the behavior of the Lasso estimators if the irrepresentable condition is relaxed. Even though the Lasso cannot recover the correct sparsity pattern, we show that the estimator is still consistent in the !2-norm sense for fixed designs under conditions on (a) the number sn of non-zero components of the vector ''n and (b) the minimal singular values of the design matrices that are induced by selecting of order sn variables. The results are extended to vectors '' in weak !q-balls with $0 < q < 1$. Our results imply that, with high probability, all important variables are selected. The set of selected variables is a useful (meaningful) reduction on the original set of variables (pn > n). Finally, our results are illustrated with the detection of closely adjacent frequencies, a problem encountered in astrophysics.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**31**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

1 Introduction

The Lasso was introduced by Tibshirani (1996) and has since been proven to be very popular and well studied (Knight and Fu, 2000; Zhao and Yu, 2006; Zou, 2005; Wainwright, 2006). Some reasons for the popularity might be that the entire regularization path of the Lasso can be computed efficiently (Osborne et al., 2000; Efron et al., 2004), that Lasso is able to handle more predictor variables than samples and produces sparse models which are easy to interpret. Several extensions and variations have been proposed (Yuan and Lin, 2005; Zhao and Yu, 2004; Zou, 2005; Meier et al., 2006; Candes and Tao, 2005b).

1.1 Lasso-type estimation

The Lasso estimator, as introduced by (Tibshirani, 1996), is given by

$$\hat{\beta}^\lambda = \operatorname{argmin}_\beta \|Y - X\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_1}, \tag{1}$$

where $X = (X_1, \ldots, X_p)$ is the $n \times p$ matrix whose columns consist of the n-dimensional fixed predictor variables X_k, $k = 1, \ldots, p$. The vector Y contains the n-dimensional set of real-valued observations of the response variable.

The distribution of Lasso-type estimators has been studied in Knight and Fu (2000). Variable selection and prediction properties of the Lasso have been studied extensively for high dimensional data with $p \gg n$, a frequently encountered challenge in modern statistical applications. Some studies (e.g. Greenshtein and Ritov, 2004; van de Geer, 2006) have focused mainly on the behavior of prediction loss. Much recent work aims at understanding the Lasso estimates from the point of view of model selection, including Meinshausen and Bühlmann (2006), Donoho et al. (2006), Zhao and Yu (2006), Candes and Tao (2005b) and Zou (2005). For the Lasso estimates to be close to the model selection estimates when the data dimensions grow, all the aforementioned papers assumed a sparse model and used various conditions that state that the irrelevant variables are not too correlated with the relevant ones. Incoherence is the terminology used in the deterministic setting of Donoho et al. (2006) and "irrespresentability" is used in the stochastic setting (linear model) of Zhao and Yu (2006). Here we focus exclusively on the properties of the estimate of the coefficient vector under squared error loss and try to understand the behavior of the estimate under a relaxed *irrepresentable condition* (hence we are in the stochastic or linear model setting). The aim is to see whether the Lasso still gives meaningful models in this case.

More discussions on the connections with other works will be covered in Section 1.5 after notions are introduced to state explicitly what the irrepresentable condition is so that the discussions are clearer.

1.2 Linear Model

We assume a linear model for the observations of the response variable $Y = (Y_1, \ldots, Y_n)$,

$$Y = X\beta + \varepsilon, \tag{2}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is a vector containing independently and identically distributed noise with $E(\varepsilon_i) = 0$. When there is a question of nonidentifiability for β when $p > n$, we define β as

$$\beta = \mathrm{argmin}_{\{\beta:\; EY=X\beta\}} \|\beta\|_{\ell_1}. \tag{3}$$

The aim is to recover the vector β as well as possible from noisy observations Y. For the equivalence between ℓ_1- and ℓ_0-sparse solutions see for example Gribonval and Nielsen (2003); Donoho and Elad (2003); Donoho (2006).

1.3 Recovery of the sparsity pattern and the irrepresentable condition

There is empirical evidence that many signals in high-dimensional spaces allow for a sparse representation. As an example, wavelet coefficients of images often exhibit exponential decay, and a relatively small subset of all wavelet coefficients allow a good approximation to the original image (Joshi et al., 1995; LoPresto et al., 1997; Mallat, 1989). For conceptual simplicity, we assume in our regression setting first that the vector β is sparse in the ℓ_0-sense and many coefficients of β are identically zero (this will later be relaxed). The corresponding variables have thus no influence on the response variable and could be safely removed. The sparsity pattern of β is understood to be the sign function of its entries, with $\mathrm{sign}(x) = 0$ if $x = 0$, $\mathrm{sign}(x) = 1$ if $x > 0$ and $\mathrm{sign}(x) = -1$ if $x < 0$. The sparsity pattern of a vector might thus look like

$$\mathrm{sign}(\beta) = (+1, -1, \quad 0, \quad 0, +1, +1, -1, +1, \quad 0, \quad 0, \ldots),$$

distinguishing whether variables have a positive, negative or no influence at all on the response variable. It is of interest whether the sparsity pattern of the Lasso estimator is a good approximation to the true sparsity pattern. If these sparsity patterns agree asymptotically, the estimator is said to be *sign consistent* (Zhao and Yu, 2006).

Definition 1 (Sign consistency) *An estimator $\hat{\beta}^\lambda$ is sign consistent if and only if*

$$P\big\{ sign(\beta) = sign(\hat{\beta}) \big\} \to 1 \qquad n \to \infty.$$

It was shown independently in Zhao and Yu (2006), Zou (2005) in the linear model case and Meinshausen and Bühlmann (2006) in Gaussian graphical model setting that *sign consistency* requires a condition on the design matrix. The assumption was termed the *irrepresentable condition* in Zhao and Yu (2006). Let $C = n^{-1}X^T X$. The dependence on n is neglected notationally.

Definition 2 (Irrepresentable condition) *Let $K = \{k : \beta_k \neq 0\}$ be the set of relevant variables and let $N = \{1, \ldots, p\} \setminus K$ be the set of noise variables. The sub-matrix C_{HK} is understood as the matrix obtained from C by keeping rows with index in the set H and columns with index in K. The irrepresentable condition is fulfilled if it holds element-wise that*

$$|C_{NK} C_{KK}^{-1} \, sign(\beta_K)| < 1.$$

In Zhao and Yu (2006), an additional *strong irrepresentable condition* is defined which requires that the above elements are not merely smaller than 1 but are uniformly bounded away from 1. Zhao and Yu (2006), Zou (2005) and Meinshausen and Bühlmann (2006) show that the Lasso is sign consistent only if the *irrepresentable condition* holds.

Proposition 1 (Sign consistency only under irrepresentable condition) *Let Assumptions 1-4 in Section 2.2 be satisfied. Assume that the irrepresentable condition is not fulfilled. Then there exists no sequence $\lambda = \lambda_n$ such that the estimator $\hat{\beta}^{\lambda_n}$ is sign consistent.*

In practice, it might be difficult to verify whether the condition is fulfilled. This led various authors to propose interesting extensions to the Lasso (Zhang and Lu, 2006; Zou, 2005; Meinshausen, 2006). Before giving up on the Lasso altogether, however, we want to examine in this paper in what sense the original Lasso procedure still gives sensible results, even if the *irrepresentable condition* is not fulfilled.

1.4 ℓ_2-consistency

The aforementioned studies showed that if the *irrepresentable condition* is not fulfilled, the Lasso cannot select the correct sparsity pattern. In this paper we show that the Lasso selects in these cases the non-zero entries of β *and some not-too-many additional* zero entries of β under relaxed conditions than the irrepresentable condition. The non-zero entries of β are in any case included in the selected model. Moreover, the size of the estimated coefficients allows to separate the few truly zero and the many non-zero coefficients. However, it is worth noting that in the extreme cases when the variables are linearly dependent, these relaxed conditions will be violated as well. In these situations, it is not sensible to use the ℓ_2-metric on β to assess Lasso. Other metrics are to be investigated in our future research.

Our main result shows the ℓ_2-consistency of the Lasso, even if the *irrepresentable condition* is relaxed. To be precise, an estimator is said to be ℓ_2-consistent if

$$\|\hat{\beta} - \beta\|_{\ell_2} \to 0 \qquad n \to \infty. \tag{4}$$

Convergence rates will also be derived. An ℓ_2-consistent estimator is attractive, as important variables are chosen with high probability and falsely chosen variables have very small coefficients. The bottom line will be that even if the sparsity pattern of β cannot be recovered by the Lasso, we can still obtain a good approximation.

1.5 Related work.

Here we discuss further the existing works on Lasso mentioned earlier on. Prediction loss for high-dimensional regression with Lipschitz loss functions under an ℓ_1-penalty is examined in van de Geer (2006). Also bounds for the ℓ_1-distance between the vector β and its Lasso estimate are derived. Similar interesting results as in van de Geer (2006) are obtained for random designs and squared error loss in Bunea et al. (2006b), both in terms of prediction loss and ℓ_1-metric. A difference with the model selection results in these papers is that we are able to obtain ℓ_2-consistency for fixed designs even if the sparsity s_n (the number of non-zero coefficients) is growing almost as fast as n, while the previous results need $s_n = o(\sqrt{n})$, partly because they require a data-independent choice of the penalty parameter. The previous study of (Bunea et al., 2006a) considers fixed designs (as we do), and obtains very nice results, albeit limited to the setting $p \leq n$, while we are interested in the high-dimensional case where the number p of predictor variables is possibly very much larger than the sample size n.

Moreover, we would like to compare the results of this manuscript briefly with results in Donoho (2004) and Candes and Tao (2005b). These papers derive bounds on the ℓ_2-norm distance between β and $\hat{\beta}$ for ℓ_1-norm constrained estimators. In Donoho (2004) the design is random and the random predictor variables are assumed to be independent. The results are thus not directly comparable to the results derived here for general fixed designs. Nevertheless, results in Meinshausen and Bühlmann (2006) suggest that the *irrepresentable condition* is with high probability fulfilled for independently normal distributed predictor variables. The results in Donoho (2004) can thus not directly be used to study the behavior of the Lasso under a violated *irrepresentable condition*, which is our goal in the current manuscript.

Candes and Tao (2005b) study the properties of the so-called "Dantzig selector", which is very similar to the Lasso, and derive remarkably sharp bounds on the ℓ_2-distance between the vector β and the proposed estimator $\hat{\beta}$. The results are derived under the condition of a *Uniform Uncertainty Principle* (UUP), which was introduced in Candes and Tao (2005a). The UUP is a relaxation of the *irrepresentable condition* and is very similar to our assumptions on sparse eigenvalues in this manuscript. It would be of interest to study the connection between the Lasso and "Dantzig selector" further, as the solutions share many similarities.

2 Main assumptions and results

First, we introduce the notion of *sparse eigenvalues*, which will play a crucial role in providing bounds for the convergence rates of the Lasso estimator. Thereafter, the assumptions are explained in detail and the main results are given.

2.1 Sparse eigenvalues

The notion of *sparse eigenvalues* is not new and has been used before (Donoho, 2006); we merely intend to fixate notation. The *m-sparse minimal eigenvalue* of a matrix is the minimal eigenvalue of any $m \times m$-dimensional submatrix.

Definition 3 *The m-sparse minimal eigenvalue and m-sparse maximal eigenvalue of C are defined as*

$$\phi_{\min}(m) = \min_{\beta:\|\beta\|_{\ell_0} \leq m} \frac{\beta^T C \beta}{\beta^T \beta}, \qquad and \qquad \phi_{\max}(m) = \max_{\beta:\|\beta\|_{\ell_0} \leq m} \frac{\beta^T C \beta}{\beta^T \beta}. \tag{5}$$

The minimal eigenvalue of the unrestricted matrix C is equivalent to $\phi_{\min}(p)$. If the number of predictor variables p is larger than sample size, $p > n$, this eigenvalue is zero, as $\phi_{\min}(m) = 0$ for any $m > n$.

A crucial factor contributing to the convergence of the Lasso estimator is the behavior of the smallest *m-sparse eigenvalue*, where the number m of variables over which the minimal eigenvalues is computed is roughly identical to the sparsity s of the true underlying vector.

2.2 Assumptions for high-dimensional data

We make some assumptions to prove the main result for high-dimensional data. We understand the term "high-dimensional" to imply here and in the following that we have potentially many more predictor variables than samples, $p_n \gg n$. While we are mainly interested in the $p_n > n$ case, the results are also relevant for $p_n \leq n$ scenarios. First, a convenient technical assumption.

Assumption 1 *The predictor variables are normalized, $\|X_k\|_{\ell_2}^2 = n$ for all k, n and* $\max_{k \in \mathbb{N}} \|X_k\|_{\ell_\infty} < \infty$.

As predictor variables are normalized in practice anyway, the first part of the assumption is not very restrictive. The second part is mostly a technical assumption and simplifies exposition.

Assumption 2 *The noise satisfies $E(\exp|\varepsilon_i|) < \infty$ and $E(\varepsilon_i^2) = \sigma^2$ for some $\sigma^2 > 0$.*

The assumption of exponential tail bounds for the noise is fairly standard and certainly covers the case of Gaussian errors.

Assumption 3 *There exists some $\sigma_y^2 < \infty$ such that $E(Y_i^2) \leq \sigma_y^2$ for all $i \in \mathbb{N}$.*

This assumption is equivalent to a re-scaling of the coefficient vectors β so that the signal-to-noise ratio stays approximately constant for all values of n in our triangular setup.

Assumption 4 *For all n, the maximal eigenvalue $\phi_{\max}(\min\{n, p\})$ for any selection of $\min\{n, p\}$ columns is bounded from above by some finite value. The minimal eigenvalue for any selection of $\min\{n, p\}$ columns is strictly positive, $\phi_{\min}(\min\{n, p\}) > 0$.*

Both parts of the assumption could be relaxed at the cost of increased notational complexity. It might be interesting to check if the assumptions are reasonable for a random design matrix. The latter part of Assumption 4 is fulfilled with probability 1 if the distribution of the random predictor variables is non-singular. Consider now the first part of the assumption about a bounded maximal eigenvalue. To be specific, assume multivariate normal predictors. If the maximal eigenvalue of the population covariance matrix, which is induced by selecting n variables, is bounded from above by an arbitrarily large constant, it follows by Theorem 2.13 Davidson and Szarek (2001) or Lemma A3.1 in Paul (2006) that the condition number of the induced sample covariance matrix observes a Gaussian tail bound. Using an entropy bound for the possible number of subsets when choosing n out of p variables, Assumption 4 is thus fulfilled with probability converging to 1 for $n \to \infty$ as long as $\log p_n = o(n^\kappa)$ for some $\kappa < 1$, and is thus maybe not overly restrictive.

2.3   Incoherent designs

As apparent from the interesting discussion in Candes and Tao (2005b), one cannot allow arbitrarily large "coherence" between variables if one still hopes to recover the correct sparsity pattern. Assume that there are two vectors β and $\tilde{\beta}$ so that the signal can be represented by either vector $X\beta = X\tilde{\beta}$ and both vectors are equally sparse, say $\|\beta\|_{\ell_0} = \|\tilde{\beta}\|_0 = s$ and are not identical. We have no hope of distinguishing between β and $\tilde{\beta}$ in such a case: if indeed $X\beta = X\tilde{\beta}$ and β and $\tilde{\beta}$ are not identical, it follows that the minimal sparse eigenvalue $\phi_{\min}(2s) = 0$ vanishes as $X(\beta - \tilde{\beta}) = 0$ and $\|\beta - \tilde{\beta}\|_{\ell_0} \le 2s$. The sparse minimal eigenvalue of a selection of order s variables indicates thus if we have any hope of recovering the true sparse underlying vector from noisy observations. A design is called m_n-*incoherent* in the following if the minimal eigenvalue of a collection of m_n variables is bounded from below by a constant.

Definition 4 (m_n-incoherent designs) *Let m_n be a sequence with $m_n = o(n)$ for $n \to \infty$. A design is called incoherent for m_n if the minimal eigenvalue of a collection of m_n variables is bounded from below, that is if*

$$\liminf_{n \to \infty} \phi_{\min}(m_n) > 0. \tag{6}$$

Our main result will require a $s_n \log n$-*incoherent design*. Most of the previous results on Lasso and related ℓ_1-constrained estimators have used similar, if slightly stronger, concepts. Donoho and Huo (2001) defined the *mutual coherence M* between two orthonormal basis as the maximal absolute value of the inner product of two elements in the two orthonormal

basis. One could extend this definition to arbitrary dictionaries where basis elements are scaled to have unit norm. Under the common assumption $M = O(1/n)$, the design is certainly *incoherent* in the meaning above, as the eigenvalue of any selection of order s_n variables, with $s_n = o(n)$, will be bounded from below by a constant for sufficiently large values of n. The notion of *incoherence* above covers thus a wider spectrum.

Candes and Tao (2005b) use a *Uniform Uncertainty Principle (UUP)* to discuss the convergence of the so-called *Dantzig selector*. The *UUP* can only be fulfilled if the minimal eigenvalue of a selection of s_n variables is bounded from below by a constant, where s_n is again the number of non-zero coefficients of β. In the original version of, a necessary condition for *(UUP)* is

$$\phi_{\min}(s_n) + \phi_{\min}(2s_n) + \phi_{\min}(3s_n) > 2.$$

In some sense, this requirement is weaker than $s_n \log n$-incoherent design as the minimal eigenvalues are calculated over maximally $3s_n$ instead of $s_n \log n$ variables. In another sense, $s_n \log n$-incoherent design is weaker as the eigenvalue can be bounded from below by an arbitrarily small constant.

Incoherent designs and the irrepresentable condition. One might ask in what sense the notion of *incoherent designs* is more general than the *irrepresentable condition*. At first, it might seem like we are simply replacing the strict condition of *irrepresentable condition* by a similarly strong condition on the design matrix.

Consider first the classical case of a fixed number p of variables. If the covariance matrix $C = C_n$ is converging to a positive definite matrix for large sample sizes, the design is automatically *incoherent*. On the other hand, it is easy to violate the *irrepresentable condition* in this case; for examples see Zou (2005).

The notion of *incoherent* designs is only a real restriction in the high-dimensional case with $p_n > n$. Even then, it is clear that the notion of *incoherence* is a relaxation from *irrepresentable condition*, as the *irrepresentable condition* can easily be violated even though all sparse eigenvalues are bounded well away from zero.

2.4   Main result for high-dimensional data $(p_n > n)$

Before we state our main result, we would like to use and explain the concept of *active variables* (Osborne et al., 2000; Efron et al., 2004) so that the penalty parameter has a useful interpretation as an upper bound on the number of active variables.

Active variables. Let G^λ be the p-dimensional gradient vector with respect to β of the squared error loss, $G^\lambda = (Y - X\hat{\beta}^\lambda)^T X$, where $\hat{\beta}^\lambda$ is the Lasso estimator. It follows by the

KKT conditions or, alternatively, results in Osborne et al. (2000) or Efron et al. (2004) that the maximum of the absolute values of the components $G^\lambda = (G_1^\lambda, \ldots, G_p^\lambda)$ is bounded by λ,

$$\max_{1 \leq k \leq p} |G_k^\lambda| \leq \lambda.$$

We call variables with maximal absolute value of the gradient *active variables*. The set of active variables is denoted by

$$\mathcal{A}_\lambda := \{k : |G_k^\lambda| = \lambda\}. \tag{7}$$

The number of selected variables (variables with a non-zero coefficient) is at most as large as the number of *active variables*, as any variable with a non-zero estimated coefficient has to be an *active variable* (Osborne et al., 2000). In Lemma 2, we derive an upper probabilistic bound on the number of active variables when setting the penalty parameter to λ. Set

$$m_\lambda := \sigma_y^2 \phi_{\max} \frac{n^2}{\lambda^2}, \tag{8}$$

where $\phi_{\max} = \phi_{\max}(\min\{n, p\})$ is the bounded maximal eigenvalue of a selection of $\min\{n, p\}$ variables. Let λ_n be a sequence of penalty parameters. Then, with probability converging to 1 for $n \to \infty$, we have that $|\mathcal{A}_{\lambda_n}| \leq m_{\lambda_n}$. Instead of the penalty parameter λ, we will often use the equivalent value of m_λ, as it offers in our opinion the better intuition.

Main result. We will discuss the implications of the theorem after the proof, as its interpretation might not be inaccessible at first.

Theorem 1 (Convergence in ℓ_2-norm) *Let Assumptions 1-4 be satisfied and assume the $s_n \log n$-incoherent design condition (6). Let m_{λ_n} be the bound (8) on the number of active variables under penalty parameter λ_n. The ℓ_2-norm of the error is then bounded for $n \to \infty$ as*

$$\|\beta - \hat{\beta}^{\lambda_n}\|_{\ell_2}^2 \leq O_p\Big(\frac{\log p_n}{n} \frac{m_{\lambda_n}}{\phi_{\min}^2(m_{\lambda_n})}\Big) + O\Big(\frac{s_n}{m_{\lambda_n}}\Big). \tag{9}$$

A proof is given in Section 3.

Remark 1 It might be of interest to compare the results for variance and bias with equivalent results for orthogonal designs. In the case of orthogonal designs, each OLS-coefficient $\hat{\beta}_k^0$, $k = 1, \ldots, p$ is soft-thresholded by the quantity $n^{-1}\lambda_n$ to get $\hat{\beta}_k^\lambda$. The squared bias of coefficients with $|\beta_k| \gg n^{-1}\lambda_n$ is thus $n^{-2}\lambda_n^2$ (under the condition that n is sufficiently large, so that $|\hat{\beta}_k^0| \geq n^{-1}\lambda_n$ with high probability). The total squared bias is thus s_n/m_{λ_n}, which is identical to the order we derive for incoherent designs so the bound cannot be improved in general.

The variance part can also be compared to the variance of an estimator for orthogonal designs, which is proportional to m_{λ_n}/n, the number of selected parameters divided by the

sample size. In our result, we get an additional $\log p_n$ factor, which stems from the fact that the subset of m_{λ_n} variables can be chosen among p_n variables. An additional factor of the reciprocal of $\phi^2_{\min}(m_{\lambda_n})$ adjusts for correlated designs.

Remark 2 It can be seen from the proofs that non-asymptotic bounds can be obtained with essentially the same results. As the constants of the non-asymptotic bounds are not very tight, we choose to present only the asymptotic result for clarity of exposition.

Organization of the remainder. The main question of concern for us is: under what circumstances can we find a sequence m_{λ_n} such that both the variance and bias term vanish asymptotically? If such a sequence exists, then we know that there is a sequence of penalty parameters λ_n so that

$$\|\beta - \hat{\beta}^{\lambda_n}\|^2_{\ell_2} \to 0 \qquad n \to \infty.$$

Sufficient conditions for ℓ_2-consistency for ℓ_0-sparse vectors are derived in the following section. Thereafter results will be extended to vectors in weak ℓ_q-balls. We will define the notion of *effective sparsity* and show that there is a one-to-one correspondence between the results for ℓ_0-sparse vectors and vectors in weak ℓ_q-balls.

Lastly, a major implication of the results is shown, namely that the Lasso can be tuned to reliably pick all important variables if selecting a small subset of the total number of variables. As already know, some unimportant variables will unfortunately also be included in this set, but can be removed in a second stage.

2.5 ℓ_2-consistency

We can immediately derive sufficient conditions for ℓ_2-consistency in the sense of (4), asking under what circumstances there exists a penalty parameter sequence λ_n so that $\|\beta - \hat{\beta}^{\lambda_n}\|_{\ell_2}$ converges in probability to 0 for large values of the sample size n. The following corollary can be derived from Theorem 1 by choosing $m_{\lambda_n} = s_n \log n$, using the incoherence assumption (6).

Corollary 1 (ℓ_2-consistency) *Let the assumptions of Theorem 1 be satisfied. The Lasso estimator is ℓ_2-consistent under the condition that*

$$s_n \, \log p_n \left(\frac{\log n}{n} \right) \to 0 \qquad n \to \infty. \tag{10}$$

The result allows thus for the number of relevant variables s_n to grow almost as fast as the number of samples n if p_n is not exponential in n, while still enjoying ℓ_2-consistency.

Remark 3 To achieve the most general result for ℓ_2-consistency, the rate λ_n of the penalty parameter has to depend on the unknown sparsity s_n. The results offer thus not so much help in picking the correct penalty parameter, but merely states that somewhere along the

solutions paths (when varying λ) there is a solution close to the true vector. How to choose the penalty parameter in a data driven way is further research.

Under less general circumstances, we can achieve ℓ_2-consistency with a fixed penalty parameter sequence that does not depend on the unknwon smoothness. Specifically, if limiting the growth rate of s_n, p_n to be $s_n \ll n^{\rho_1}$ and $n/\log p_n \gg n^{\rho_2}$ for some $0 \le \rho_1 < \rho_2 < 1$, then any sequence $m_{\lambda_n} \asymp n^\rho$ with $\rho_1 < \rho < \rho_2$ achieves ℓ_2-consistency, irrespective of the actual sparsity, as long as the stronger incoherence assumption $\liminf_{n\to\infty} \phi_{\min}(m_{\lambda_n}) > 0$ is fulfilled.

2.6 Some results for weak ℓ_q-balls

So far, we have been assuming that the vector β is sparse in an ℓ_0-sense, with most entries of β being identically zero. This is a conceptually simple assumption. It is easy to formulate the problem and understand the results for the ℓ_0-sparse setting. In the end, however, it might be overly simplistic to assume that most of the entries are identically zero. It is perhaps more interesting to assume that most of the entries are very small, as is the case for wavelet coefficients of natural images (Joshi et al., 1995; LoPresto et al., 1997; Mallat, 1989). We can for example consider the case that the vector β lies in a weak ℓ_q-ball with $0 < q < 2$. Let $|\beta_{(1)}| \ge |\beta_{(2)}| \ge \ldots \ge |\beta_{(p)}|$ be the ordered entries of β. The vector β lies in a weak ℓ_1-ball if there exists a constant $s_{q,n} > 0$ such that

$$\forall 1 \le k \le p : \qquad |\beta_{(k)}| \le s_{q,n} \, k^{-1/q}. \tag{11}$$

If a vector β has a ℓ_q-(quasi-)norm $\|\beta\|_{\ell_q}$, then it also lies in a weak ℓ_q-ball with the sparsity $s_{q,n}$, that is

$$s_{q,n} \le \|\beta\|_{\ell_q}.$$

In the ℓ_q-sparse setting, it does not make sense trying to recover the correct sparsity pattern, as all coefficients are in general different from zero. We can, however, ask if the most important coefficients are recovered, neglecting coefficients with very small absolute value. Consider the case $0 < q < 1$, where coefficients decay faster than $1/k$. As can be seen in the following, the bound on the ℓ_2-distance between β and its estimate $\hat{\beta}^\lambda$ is very similar to the ℓ_0-sparse setting.

Effective Sparsity. There is a simple connection between the results for ℓ_0-sparse and ℓ_q-sparse vectors. Specifically, we are interested in settings where ℓ_2-consistency of the Lasso estimator can be achieved. If we define the effective sparsity as $s_{q,n}$ raised to the power of $2q/(2-q)$, the results of the ℓ_0-sparse setting are directly applicable to the ℓ_q-sparse setting with $0 < q < 1$.

Definition 5 (Effective sparsity) *The effective sparsity s_n^{eff} of a vector β in a weak ℓ_q-ball with sparsity $s_{q,n}$ is defined as*

$$s_n^{eff} = s_{q,n}^{\frac{2q}{2-q}}. \tag{12}$$

To motivate the notion of *effective sparisty*, suppose that the decay of the absolute value of the components of β is fast. A good approximation to β in the ℓ_2-sense can then be obtained by retaining just a few large components of β. Assume that the entries of β are ordered so that $|\beta_1| \geq |\beta_2| \geq \ldots \geq |\beta_p|$. Let $\tilde{\beta}^d$ be the approximation that retains only the d largest components,

$$\tilde{\beta}_k^d = \begin{cases} \beta_k & k \leq d \\ 0 & d > s \end{cases}$$

The *effective sparsity* measures the minimal number $d = d_n$ of non-zero components necessary to obtain an approximation $\tilde{\beta}^{d_n}$ of β that satisfies

$$\|\tilde{\beta}^{d_n} - \beta\|_{\ell_2} \to 0 \qquad n \to \infty. \tag{13}$$

To be precise, let $\mathcal{B}$ be the set $\mathcal{B} = \{\beta : \|\beta\|_{\ell_q} \leq s_{q,n} \text{ for all } n\}$. Then, for any sequence d_n which satisfies (13) for every vector $\beta \in \mathcal{B}$, the number of retained coefficients d_n needs to be at least of the same order as the *effective sparsity*, that is $\liminf_{n\to\infty} d_n/s_n^{eff} > 0$. On the other hand, retaining exactly s_n^{eff} components satisfies (13) for every vector $\beta \in \mathcal{B}$. A proof of this is straightforward but omitted here. The notion of *effective sparsity* is thus inherent to the nature of the problem.

The definition of *effective sparsity* will be helpful in the following. Suppose we want to see whether ℓ_2-consistency can be achieved for vectors in a weak ℓ_q-ball with sparsity $s_{q,n}$. The *effective sparsity* of this setting can then be calculated according to (12). We can then look at the ℓ_0-sparse vectors, where the number of non-zero entries is set to the *effective sparsity* of the original problem. If ℓ_2-consistency can be achieved for the ℓ_0-sparse setting, it can also be achieved for the ℓ_q-sparse setting. With the notion of *effective sparsity*, a bound on the ℓ_2-distance between the Lasso estimator and the true vector can be derived.

Theorem 2 *Let the assumptions of Theorem 1 be fulfilled, except that the vector β is only assumed to be in a weak ℓ_q-ball for some $0 < q < 1$. Let again m_{λ_n} be the bound on the number of active variables (8) and assume m_n-incoherent design. Then*

$$\|\beta - \hat{\beta}^{\lambda_n}\|_{\ell_2}^2 = O_p\left(\frac{\log p_n}{n} m_{\lambda_n}\right) + O\left\{\left(\frac{s_n^{eff}}{m_{\lambda_n}}\right)^{1-q/2}\right\}.$$

A proof is given in Section 4.

The place of the sparsity measure s_n, the number of non-zero elements, is now taken by the effective sparsity s_n^{eff}. The implications of Theorem 2 are similar to those of ℓ_0-sparse vectors. The Lasso is thus able to recover not only ℓ_0-sparse vectors but also vectors which are sparse in the sense of lying in a weak ℓ_q-ball for some small value of q.

2.7 Sign consistency with two-step procedures

The results show that the Lasso estimator can be made *sign consistent* in a two-step procedure even if the *irrepresentable condition* is relaxed but under the assumption that non-zero coefficients of β are "sufficiently" large. One possibility is hard-thresholding of the obtained coefficients, neglecting variables with very small coefficients. This effect has already been observed empirically in (Valdés-Sosa et al., 2005). Other possibilities include soft-thresholding and relaxation methods such as the Gauss-Dantzig selector (Candes and Tao, 2005b) or the Relaxed Lasso (Meinshausen, 2006) with an additional thresholding step.

We start with a corollary that follows directly from Theorem 1, stating that important variables are chosen with high probability. Let L be the subset of large coefficients and Z be the subset of zero coefficients,

$$
\begin{aligned}
L &:= \{k : \beta_k^4 \gg \frac{s_n \log p_n}{n}\}, \\
Z &:= \{k : \beta_k = 0\}
\end{aligned}
$$

where $a_n \gg b_n$ is again meant to imply that $a_n/b_n \to \infty$ for $n \to \infty$. The following corollary states that the Lasso can distinguish between variables in L and Z.

Corollary 2 *Let the assumptions of Theorem 2 be satisfied and assume $\sqrt{s_n n}$-incoherent design. There exists a penalty sequence λ_n such that, with probability converging to 1 for $n \to \infty$, the absolute value of coefficients in L is larger than the absolute value of any coefficient in Z,*

$$
\forall k_l \in L, k_z \in Z : \qquad |\hat{\beta}_{k_l}^{\lambda_n}| > |\hat{\beta}_{k_z}^{\lambda_n}|. \tag{14}
$$

Proof. The proof follows from the results of Theorem 1. Event (14) is fulfilled if $\|\hat{\beta}^{\lambda_n} - \beta\|_\infty \le \min_{k \in L} |\beta_k|$. Choosing a penalty sequence λ_n with $m_{\lambda_n}^2 = n s_n \log^{-1} p_n$ yields with Theorem 1 that $\|\hat{\beta}^{\lambda_n} - \beta\|_{\ell_2}^4 \le O_p(n^{-1} s_n \log p_n)$. The bound on the ℓ_2-distance gives trivially the identical bound on the ℓ_∞-distance between $\hat{\beta}^{\lambda_n}$ and β. Furthermore, by definition of the set L, $\min_{k \in L} |\beta_k| \gg (n^{-1} s_n \log p_n)^{1/4}$, which comples the proof. $\square$

Remark 4 The corollary implies that variables with sufficiently large regression coefficients are chosen with very large probability by the Lasso, that is for $n \to \infty$,

$$
P(\forall k \in L : \hat{\beta}_k^{\lambda_n} \ne 0) \to 1.
$$

As also some additional unwanted variables are chosen (which cannot be avoided if the *irrepresentable condition* is violated), the result implies that the Lasso is successful in narrowing down the choice of $p_n \gg n$ variables to a subset of variables with cardinality much smaller than n (at least of smaller order than $\sqrt{n s_n}$). All important variables are with large

13

probability in this much smaller Lasso-selected subset. A two-step procedure would try to filter out those important variables from the selected subset. Consistent variable selection could for example be achieved by simple thresholding of small coefficients in the initial Lasso estimator.

Remark 5 It is apparent that the large bias of the Lasso estimator allows only for a slow rate of decay of coefficients in the set L. To alleviate this problem, one could first reduce the bias of the selected coefficients and apply thresholding after this relaxation step. Even though we view this bias-reduction step as important, we refrain from giving more details due to space constraints.

In conclusion, even though one cannot achieve *sign consistency* in general with just a single Lasso estimation, it can often be achieved in a two-stage procedure.

3 Proof of Theorem 1

The first term on the right hand side of (9) is a variance-type and the second term represent a bias-type contribution. Let β_n^λ be the estimator under the absence of noise, that is $\beta^\lambda = \hat{\beta}^{\lambda,0}$, where $\hat{\beta}^{\lambda,\xi}$ is defined as in (30). The ℓ_2-distance can then be bounded by $\|\hat{\beta}^\lambda - \beta\|_{\ell_2}^2 \leq 2\|\hat{\beta}^\lambda - \beta^\lambda\|_{\ell_2}^2 + 2\|\beta^\lambda - \beta\|_{\ell_2}^2$. The first term on the right hand side represents the variance of the estimation, while the second term represents the bias. The bound on the variance term follows by Lemma 6. The bias contribution follows directly from Lemma 1. □

3.1 Part I of Proof: Bias

Let K be the set of non-zero elements of β, that is $K = \{k : \beta_k \neq 0\}$. The cardinality of K is again denoted by $s = s_n$. For the following, let β^λ be the estimator $\hat{\beta}^\lambda$ under the absence of noise, $\sigma = 0$. The solution β^λ can, for each value of λ, be written as $\beta^\lambda = \beta + \gamma^\lambda$, where

$$\gamma^\lambda = \text{argmin}_{\zeta \in \mathbb{R}^p} \, f(\zeta), \tag{15}$$

where the function $f(\zeta)$ is given by

$$f(\zeta) = n\zeta^T C \zeta + \lambda \sum_{k \in K^c} |\zeta_k| + \lambda \sum_{k \in K} \left(|\beta_k + \zeta_k| - |\beta_k| \right). \tag{16}$$

The vector γ^λ is the bias of the Lasso estimator. We derive first a bound on the ℓ_2-norm of γ^λ.

Lemma 1 *Let $\kappa > 0$ be the minimal eigenvalue with $\liminf_{n \to \infty} \phi_{\min}(s_n \log s_n) > \kappa$. The ℓ_2-norm of γ^λ, as defined in (15), is bounded for sufficiently large values of n by*

$$\|\gamma^\lambda\|_{\ell_2} \leq \frac{\lambda}{n} \frac{\sqrt{s_n}}{\kappa}.$$

Proof. We write in the following γ instead of γ^λ for notational simplicity. Let $\gamma(K)$ be the vector with coefficients $\gamma_k(K) = \gamma_k 1\{k \in K\}$, that is $\gamma(K)$ is the bias of the truly non-zero coefficients. Analogously, let $\gamma(K^c)$ be the bias of the truly zero coefficients with $\gamma_k(K^c) = \gamma_k 1\{k \notin K\}$. Clearly, $\gamma = \gamma(K) + \gamma(K^c)$. The value of the function $f(\zeta)$, as defined in (16), is 0 if setting $\zeta = 0$. For the true solution γ^λ, it follows hence that $f(\gamma^\lambda) \leq 0$. Hence, using that $\zeta^T C \zeta \geq 0$ for any ζ,

$$\|\gamma(K^c)\|_{\ell_1} = \sum_{k \in K^c} |\zeta_k| \leq \big| \sum_{k \in K} \big(|\beta_k + \zeta_k| - |\beta_k| \big) \big| \leq \|\gamma(K)\|_{\ell_1}. \tag{17}$$

As $\|\gamma(K)\|_{\ell_0} \leq s_n$, it follows that $\|\gamma(K)\|_{\ell_1} \leq \sqrt{s_n}\|\gamma(K)\|_{\ell_2} \leq \sqrt{s_n}\|\gamma\|_{\ell_2}$ and hence, using (17),

$$\|\gamma\|_{\ell_1} \leq 2\sqrt{s_n}\|\gamma\|_{\ell_2}. \tag{18}$$

This result will be used further below. We use now again that $f(\gamma^\lambda) \leq 0$ (as $\zeta = 0$ yields the upper bound $f(\zeta) = 0$). Using the previous result that $\|\gamma(K)\|_{\ell_1} \leq \sqrt{s_n}\|\gamma\|_{\ell_2}$, and ignoring the non-negative term $\|\gamma(K^c)\|_{\ell_1}$, it follows that

$$n\gamma^T C \gamma \leq \lambda\sqrt{s_n}\|\gamma\|_{\ell_2}. \tag{19}$$

Consider now the term $\gamma^T C \gamma$. Bounding this term from below and plugging the result into (19) will yield the desired upper bound on the ℓ_2-norm of γ. Let $|\gamma_{(1)}| \geq |\gamma_{(2)}| \geq \ldots \geq |\gamma_{(p)}|$ be the ordered entries of γ.

Let $\{u_n\}_{n \in \mathbb{N}}$ be a sequence of positive integers, to be chosen later, and define the set of the "u_n-largest coefficients" as $U = \{k : |\gamma_k| \geq |\gamma_{(u_n)}|\}$. Define analogously to above the vectors $\gamma(U)$ and $\gamma(U^c)$ by $\gamma_k(U) = \gamma_k 1\{k \in U\}$ and $\gamma_k(U^c) = \gamma_k 1\{k \notin U\}$. The quantity $\gamma^T C \gamma$ can be written as

$$\gamma^T C \gamma = \gamma(U)^T C \gamma(U) = \|a + b\|_{\ell_2}^2, \tag{20}$$

where $a := n^{-1/2} X \gamma(U)$ and $b := n^{-1/2} X \gamma(U^c)$. Then

$$\gamma^T C \gamma = a^T a + 2b^T a + b^T b \geq \|a\|_{\ell_2}^2 - 2\|a\|_{\ell_2}\|b\|_{\ell_2}. \tag{21}$$

As $\gamma(U)$ has by definition only u_n non-zero coefficients,

$$\|a\|_{\ell_2}^2 = \|\gamma(U)^T C \gamma(U)\|_{\ell_2}^2 \geq \phi_{\min}(u_n)\|\gamma(U)\|_{\ell_2}^2 = \phi_{\min}(u_n)(\|\gamma\|_{\ell_2}^2 - \|\gamma(U^c)\|_{\ell_2}^2). \tag{22}$$

As $\gamma(U^c)$ has at most n non-zero coefficients,

$$\|b\|_{\ell_2}^2 = \|\gamma(U^c)^T C \gamma(U^c)\|_{\ell_2}^2 \leq \phi_{\max}(n)\|\gamma(U^c)\|_{\ell_2}^2 \tag{23}$$

Using (22) and (23) in (21),

$$\begin{aligned}
\|a\|_{\ell_2}^2 - 2\|a\|_{\ell_2}\|b\|_{\ell_2} \geq{}& \phi_{\min}(u_n)(\|\gamma\|_{\ell_2}^2 - \|\gamma(U^c)\|_{\ell_2}^2) \\
&- 2\sqrt{\phi_{\min}(u_n)\phi_{\max}(n)}\|\gamma\|_{\ell_2}\|\gamma(U^c)\|_{\ell_2}.
\end{aligned}$$

15

and hence

$$\gamma^T C \gamma \geq \phi_{\min}(u_n) \|\gamma\|_{\ell_2}^2 \left(1 - 2 \sqrt{\frac{\phi_{\max}(n)}{\phi_{\min}(u_n)}} \frac{\|\gamma(U^c)\|_{\ell_2}}{\|\gamma\|_{\ell_2}} - \frac{\|\gamma(U^c)\|_{\ell_2}^2}{\|\gamma\|_{\ell_2}^2} \right). \tag{24}$$

Before proceeding, we need to bound the norm $\|\gamma(U^c)\|_{\ell_2}$ as a function of u_n. Assume for the moment that the ℓ_1-norm $\|\gamma\|_{\ell_1}$ is identical to some $\ell > 0$. Then it holds for every $k = 1, \ldots, p$ that $\gamma_{(k)} \leq \ell/k$. Hence,

$$\|\gamma(U^c)\|_{\ell_2}^2 \leq \|\gamma\|_{\ell_1}^2 \sum_{k=u_n+1}^{p} \frac{1}{k^2} \leq (4 s_n \|\gamma\|_{\ell_2}^2) \frac{1}{u_n}, \tag{25}$$

having used the result (18) from above that $\|\gamma\|_{\ell_1} \leq 2\sqrt{s_n}\|\gamma\|_{\ell_2}$. Plugging this result into (24),

$$\gamma^T C \gamma \geq \phi_{\min}(u_n) \|\gamma\|_{\ell_2}^2 \left(1 - 2 \sqrt{\frac{4 s_n \phi_{\max}(n)}{u_n \phi_{\min}(u_n)}} - \frac{4 s_n}{u_n} \right) \tag{26}$$

Choosing a sequence $u_n = s_n \log n$, it holds with the assumption $\liminf_{n \to \infty} \phi_{\min}(s_n \log n) > \kappa$, that for sufficiently large values of n, by the assumption of a bounded maximal eigenvalue $\phi_{\max}(n)$,

$$\gamma^T C \gamma \geq \kappa \|\gamma\|_{\ell_2}^2.$$

Using the last result together with (19), which says that $\gamma^T C \gamma \leq n^{-1} \lambda \sqrt{s_n} \|\gamma\|_{\ell_2}$, it follows that for large n,

$$\|\gamma\|_{\ell_2} \leq \frac{\lambda}{n} \frac{\sqrt{s_n}}{\kappa},$$

which completes the proof. $\qquad \square$

3.2   Part II of Proof: Variance

First, bounds for the number of selected and active variables are derived. These bounds are later used to assess the variance of the estimator under noisy observations.

A bounds on the number of active variables   A decisive part in the variance of the estimator is determined by the number of selected variables. Instead of directly bounding the number of selected variables, we derive bounds for the number of *active variables*. As any variable with a non-zero regression coefficient is also a *active variable*, these bounds lead trivially to bounds for the number of selected variables.

Let again $\mathcal{A}_\lambda$ be the set of active variables,

$$\mathcal{A}_\lambda = \{ k : |G_k^\lambda| = \lambda \}.$$

16

The number of selected variables (variables with a non-zero coefficient) is at most as large as the number of *active variables*, as any variable with a non-zero estimated coefficient has to be an *active variable* (Osborne et al., 2000).

Lemma 2 *With probability tending to 1 for $n \to \infty$, the number $|\mathcal{A}_\lambda|$ of active variables of the estimator $\hat{\beta}^\lambda$ is bounded by*

$$|\mathcal{A}_\lambda| \leq \sigma_y^2 \phi_{\max}(|\mathcal{A}_\lambda|) \frac{n^2}{\lambda^2} \leq \sigma_y^2 \phi_{\max}(\min\{n,p\}) \frac{n^2}{\lambda^2} := m_\lambda,$$

where $\phi_{\max}(|\mathcal{A}_\lambda|)$ is the bounded maximal eigenvalue of a selection of at most $|\mathcal{A}_\lambda| \leq \min\{n,p\}$ variables.

Proof. Let $R(\lambda)$ be the vector residuals, $R(\lambda) = Y - X\hat{\beta}^\lambda$. For any k in the $|\mathcal{A}_\lambda|$-dimensional space spanned by the *active variables*,

$$|G_k^\lambda| = |R^T(\lambda)X_k| = \lambda. \tag{27}$$

Let $R^{\mathcal{A}}(\lambda)$ be the projection $P^{\mathcal{A}}R(\lambda)$ of the residuals $R(\lambda)$ into the $|\mathcal{A}_\lambda|$-dimensional space spanned by the $|\mathcal{A}_\lambda|$ active variables. Then, by (27),

$$\|X_{\mathcal{A}}^T R^{\mathcal{A}}(\lambda)\|_{\ell_2}^2 = \|X_{\mathcal{A}}^T R(\lambda)\|_{\ell_2}^2 = |\mathcal{A}_\lambda|\lambda^2. \tag{28}$$

As $R^{\mathcal{A}}(\lambda)$ is the projection onto the space spanned by $|\mathcal{A}_\lambda|$ active variables, it holds that for $|\mathcal{A}_\lambda| \leq n$, with the notation $v = X_{\mathcal{A}}^T R(\lambda)$,

$$R^{\mathcal{A}}(\lambda) = X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}v,$$

and hence

$$\|R^{\mathcal{A}}(\lambda)\|_{\ell_2}^2 = v^T(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}v \geq \{n\phi_{\max}(|\mathcal{A}_\lambda|)\}^{-1}\|v\|_{\ell_2}^2.$$

Using the result (28), it follows that $\|v\|_{\ell_2}^2 \geq |\mathcal{A}_\lambda|\lambda^2$ and hence

$$\|R^{\mathcal{A}}(\lambda)\|_{\ell_2}^2 \geq \{n\phi_{\max}(|\mathcal{A}_\lambda|)\}^{-1}|\mathcal{A}_\lambda|\lambda^2.$$

The sum of squared residuals is bounded uniformly over all subsets M of the active variables by $\|R^{\mathcal{A}}(\lambda)\|_{\ell_2}^2 \leq \|Y\|_{\ell_2}^2$. By assumption 3, it holds with probability converging to 1 for $n \to \infty$, that $\|Y\|_{\ell_2}^2 \leq n\sigma_y^2$, which completes the proof. □

De-noised response We need for the following a little extension of the result above. Define for $0 < \xi < 1$ the de-noised version of the response variable,

$$Y(\xi) = X\beta + \xi\varepsilon. \tag{29}$$

We can regulate the amount of noise with the parameter ξ. For $\xi = 0$, only the signal is retained. The original observations with the full amount of noise are recovered for $\xi = 1$. Now consider for $0 \le \xi \le 1$ the estimator $\hat{\beta}^{\lambda,\xi}$,

$$\hat{\beta}^{\lambda,\xi} \;=\; \mathrm{argmin}_\beta \|Y(\xi) - X\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_1}. \tag{30}$$

The ordinary Lasso estimate is recovered under the full amount of noise so that $\hat{\beta}^{\lambda,1} = \hat{\beta}^\lambda$. Using the notation from the previous results, we can write for the estimate in the absence of noise, $\hat{\beta}^{\lambda,0} = \beta^\lambda$. The definition of the de-noised version of the Lasso estimator will be helpful for the proof as it allows to characterize the variance of the estimator.

Number of active variables for the de-noised estimator. In analogy to the gradient G^λ of the loss function, let $G^{\lambda,\xi}$ be the gradient vector with respect to β of the squared error loss when estimating the de-noised version of the observations,

$$G^{\lambda,\xi} = (Y(\xi) - X\hat{\beta}^{\lambda,\xi})^T X. \tag{31}$$

Variables are called again *active* if the absolute value of the respective gradient is equal to the maximal value λ. The *active variables* are denoted by $\mathcal{A}_{\lambda,\xi}$. The result of Lemma 2 can now easily be shown to hold uniformly over all de-noised versions of the estimator.

Lemma 3 *With probability converging to 1 for $n \to \infty$, the number $|\mathcal{A}_{\lambda,\xi}|$ of active variables of the de-noised estimator is bounded by*

$$\left|\mathcal{A}_{\lambda,\xi}\right| \le \sigma_y^2 \frac{n^2}{\lambda^2}\phi_{\max}(\left|\mathcal{A}_{\lambda,\xi}\right|.$$

Proof. The proof follows analogously to the proof of Lemma 2. In analogy to $R^{\mathcal{A}}$, let $R^{\mathcal{A},\xi}$ be the projection of the residuals $Y(\xi) - X\hat{\beta}^{\lambda,\xi}$ onto the space spanned by the *active variables*. The bound $\|R^{\mathcal{A},\xi}\|_{\ell_2} \le \sup_{0\le\xi\le1} \|Y(\xi)\|_{\ell_2}$ holds uniformly over all values of ξ with $0 \le \xi \le 1$. By assumption 3, the term $n^{-1}\sup_{0\le\xi\le1} \|Y(\xi)\|_{\ell_2}$ is, with probability converging to 1 for $n \to \infty$, bounded by σ_y^2. The proof follows then exactly like the proof of Lemma 2. $\square$

This uniform bound is used below to bound the variance of the Lasso estimator.

Variance of restricted OLS   Next, we consider the variance of the Lasso estimator as a function of the penalty parameter. Let $\hat{\theta}^M \in \mathbb{R}^p$ be for every subset $M \subseteq \{1, \ldots, p\}$ with $|M| \leq n$ the restricted OLS-estimator of the noise vector ε,

$$\hat{\theta}^M = (X_M^T X_M)^{-1} X_M^T \varepsilon. \tag{32}$$

First, we bound the ℓ_2-norm of this estimator. The result is useful for bounding the variance of the final estimator, based on the derived bound on the number of active variables.

Lemma 4 *The ℓ_2-norm of the restricted estimator $\hat{\theta}^M$ is uniformly over all sets M with $|M| \leq m_{\lambda_n}$, where m_{λ_n} is as defined in (8), bounded for $n \to \infty$ by*

$$\max_{M:|M| \leq m_{\lambda_n}} \|\hat{\theta}^M\|_{\ell_2}^2 \quad = \quad O_p\Big(\frac{\log p_n}{n} \frac{m_{\lambda_n}}{\phi_{\min}^2(m_{\lambda_n})}\Big).$$

Proof.   It follows directly from the definition of $\hat{\theta}^M$ that, for every M with $|M| \leq m_{\lambda_n}$,

$$\|\hat{\theta}^M\|_{\ell_2}^2 \leq \frac{1}{n^2 \phi_{\min}^2(m_{\lambda_n})} \|X_M^T \varepsilon\|_{\ell_2}^2. \tag{33}$$

It remains to be shown that, for $n \to \infty$,

$$\max_{M:|M| \leq m_{\lambda_n}} \|X_M^T \varepsilon\|_{\ell_2}^2 = O_p(m_{\lambda_n} n \log p_n).$$

As $E(\exp|\varepsilon_i|) < \infty$ and $\max_k \|X_k\|_{\ell_\infty} < \infty$, it follows by Bernstein's inequality that $|X_k^T \varepsilon|^2 = O_p(n)$ for every $k \leq p_n$ and due to the exponential tail bound

$$\max_{k \leq p_n} |X_k^T \varepsilon|^2 \quad = \quad O_p(n \log p_n);$$

and hence,

$$\max_{M:|M| \leq m_{\lambda_n}} \|X_M^T \varepsilon\|_{\ell_2}^2 \leq m_{\lambda_n} \max_{k \leq p_n} |X_k^T \varepsilon|^2 = O_p(n m_{\lambda_n} \log p_n).$$

Using this in conjunction with (33) completes the proof. $\square$

Variance of estimate is bounded by restricted OLS variance   We show that the variance of the Lasso estimator can be bounded by the variances of restricted OLS estimators, using bounds on the number of active variables.

Lemma 5 *If, for a fixed value of λ, the number of active variables of the de-noised estimators $\hat{\beta}^{\lambda,\xi}$ is for every $0 \leq \xi \leq 1$ bounded by m, then*

$$\|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda,1}\|_{\ell_2}^2 \quad \leq \quad \max_{M:|M| \leq m} \|\hat{\theta}^M\|_{\ell_2}^2.$$

Proof. The key in the proof is that the solution path of $\hat{\beta}^{\lambda,\xi}$, if increasing the value of ξ from 0 to 1, can be expressed piecewise in terms of the restricted OLS solution. The set $M(\xi)$ of active variables is the set with maximal absolute gradient,

$$M(\xi) = \{k : |G_k^{\lambda,\xi}| = \lambda\}.$$

Note that the estimator $\hat{\beta}^{\lambda,\xi}$ and also the gradient $G_k^{\lambda,\xi}$ are continuous functions in both λ and ξ (Efron et al., 2004). Let $0 = \xi_1 < \xi_2 < \ldots < \xi_{L+1} = 1$ be the points of discontinuity of $M(\xi)$. At these locations, variables either join the active set or are dropped from the active set.

Fix some j with $1 \leq j \leq J$. Denote by M_j the set of active variables $M(\xi)$ for any $\xi \in (\xi_j, \xi_{j+1})$. We show in the following that the solution $\hat{\beta}^{\lambda,\xi}$ is for all ξ in the interval (ξ_j, ξ_{j+1}) given by

$$\forall \xi \in (\xi_j, \xi_{j+1}): \qquad \hat{\beta}^{\lambda,\xi} = \hat{\beta}^{\lambda,\xi_j} + (\xi - \xi_j)\hat{\theta}^{M_j}, \tag{34}$$

where $\hat{\theta}^{M_j}$ is the restricted OLS estimator of noise, as defined in (32). The local effect of increased noise (larger value of ξ) on the estimator is thus to shift the coefficients of the active set of variables along the least squares direction.

Once (34) is shown, the claim follows by piecing together the piecewise linear parts and using continuity of the solution as a a function of ξ to obtain

$$
\begin{aligned}
\|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda,1}\|_{\ell_2} &\leq \sum_{j=1}^{J} \|\hat{\beta}^{\lambda,\xi_j} - \hat{\beta}^{\lambda,\xi_{j+1}}\|_{\ell_2} \\
&\leq \max_{M:|M|\leq m} \|\hat{\theta}^M\|_{\ell_2} \sum_{j=1}^{J} (\xi_{j+1} - \xi_j) = \max_{M:|M|\leq m} \|\hat{\theta}^M\|_{\ell_2}.
\end{aligned}
$$

It thus remains to show (34). A necessary and sufficient condition for $\hat{\beta}^{\lambda,\xi}$ with $\xi \in (\xi_j, \xi_{j+1})$ to be a valid solution is that for all $k \in M_j$ with non-zero coefficient $\hat{\beta}_k^{\lambda,\xi} \neq 0$, the gradient is equal to λ times the negative sign,

$$G_k^{\lambda,\xi} = -\lambda \, \text{sign}(\hat{\beta}_k^{\lambda,\xi}), \tag{35}$$

that for all variables with $k \in M_j$ with zero coefficient $\hat{\beta}_k^{\lambda,\xi} = 0$ the gradient is equal in absolute value to λ

$$|G_k^{\lambda,\xi}| = \lambda, \tag{36}$$

and for variables $k \notin M_j$ not in the active set,

$$|G_k^{\lambda,\xi}| < \lambda. \tag{37}$$

These conditions are a consequence of the requirement that the subgradient of the loss function contains 0 for a valid solution.

Note that the gradient of the active variables in M_j is unchanged if replacing $\xi \in (\xi_j, \xi_{j+1})$ by some $\xi' \in (\xi_j, \xi_{j+1})$ and replacing $\hat{\beta}^{\lambda,\xi}$ by $\hat{\beta}^{\lambda,\xi} + (\xi' - \xi)\hat{\theta}^{M_j}$. That is, for all $k \in M_j$,

$$\big(Y(\xi) - X\hat{\beta}^{\lambda,\xi}\big)^T X_k = \big\{Y(\xi') - X\big(\hat{\beta}^{\lambda,\xi} + (\xi' - \xi)\hat{\theta}^{M_j}\big)\big\}^T X_k,$$

as the difference of both sides is equal to

$$(\xi' - \xi)\big(\varepsilon - X\hat{\theta}^{M_j}\big)^T X_k,$$

and $\big(\varepsilon - X\hat{\theta}^{M_j}\big)^T X_k = 0$ for all $k \in M_j$, as $\hat{\theta}^{M_j}$ is the OLS of ε, regressed on the variables in M_j. Equalities (35) and (36) are thus fulfilled for the solution and it remains to show that (37) also holds. For sufficiently small values of $\xi' - \xi$, inequality (37) is clearly fulfilled for continuity reasons. Note that if $|\xi' - \xi|$ is large enough such that for one variable $k \notin M_j$ inequality (37) becomes an equality, then the set of active variables changes and thus either $\xi' = \xi_{j+1}$ or $\xi' = \xi_j$. We have thus shown that the solution $\hat{\beta}^{\lambda,\xi}$ can for all $\xi \in (\xi_j, \xi_{j+1})$ be written as

$$\hat{\beta}^{\lambda,\xi} = \hat{\beta}^{\lambda,\xi_j} + (\xi - \xi_j)\hat{\theta}^{M_j},$$

which proves (34) and thus completes the proof. □

Lemma 6 *Under the assumptions of Theorem 1, the variance term is bounded by*

$$\|\hat{\beta}^\lambda - \beta^\lambda\|_{\ell_2}^2 \quad = \quad O_p\Big(\frac{\log p_n}{n} \frac{m_{\lambda_n}}{\phi_{\min}^2(m_{\lambda_n})}\Big),$$

where $m_{\lambda_n} = n^2 \lambda^{-2} \sigma_y^2 \phi_{\max}(\min\{n, p\})$.

By Lemma 5 and 4, the variance can be bounded by

$$\|\hat{\beta}^\lambda - \beta^\lambda\|_{\ell_2}^2 = O_p\Big(\frac{\log p_n}{n} \frac{\tilde{m}}{\phi_{\min}^2(\tilde{m})}\Big),$$

where $\tilde{m} = \sup_{0 \leq \xi \leq 1} m^{\lambda,\xi}$ is the maximal number of active variable of the de-noised estimate (30). Using Lemma 3, the number $\tilde{m}$ of active variables is bounded, with probability converging to 1 for $n \to \infty$, by $m_{\lambda_n} = n^2 \lambda_n^{-2} \sigma_y^2 \phi_{\max}(\min\{n, p\})$, which completes the proof.

4 Proof of Theorem 2

The proof of Theorem 2 is in most parts analogous to the proof of Lemma 1 as the bound on the variance part remains unchanged. Only a bound on the ℓ_2-norm of the bias has to be recalculated. First, we derive a bound on the ℓ_1-norm of γ^λ, similar to (17). The solution β^λ can be written as $\beta + \gamma^\lambda$, where

$$\gamma^\lambda = \text{argmin}_{\zeta \in \mathbb{R}^p} f(\zeta), \tag{38}$$

and the function $f(\zeta)$ defined in (16) is now written as

$$f(\zeta) = n\zeta^T C\zeta + \lambda \sum_{k \le p} \left(|\beta_k + \zeta_k| - |\beta_k| \right).$$

In the proof of Lemma 1, K was the set of variables with truly non-zero coefficients. In the current setting, all coefficients are potentially different from zero. Let instead K be in the following the set of variables whose coefficient is among the r_n largest, where r_n is some sequence (to be chosen later) with $r_n \to \infty$ for $n \to \infty$. Assume without loss of generality that $|\beta_1| \ge |\beta_2| \ge \ldots \ge |\beta_p|$. Then,

$$K = \{k : k \le r_n\}.$$

We can use again that $f(\gamma^\lambda) \le 0$ as setting $\zeta = 0$ yields already the upper bound $f(0) = 0$. Using that C is positive semi-definite,

$$\sum_{k \in K^c} (|\beta_k + \gamma_k| - |\beta_k|) \le -\sum_{k \in K}(|\beta_k + \gamma_k| - |\beta_k|).$$

Note that on the one hand $|c+d| - |d| \le |c|$ for all $c, d \in \mathbb{R}$. On the other hand, $|c+d| - |d| \ge |c| - 2|d|$ for all $c, d \in \mathbb{R}$. Thus

$$\|\gamma(K^c)\|_{\ell_1} - 2\|\beta(K^c)\|_{\ell_1} = \sum_{k \in K^c} (|\gamma_k| - 2|\beta_k|) \le \sum_{k \in K} |\gamma_k| = \|\gamma(K)\|_{\ell_1}.$$

As β lies in a weak ℓ_q-ball with $0 < q < 1$, we have, by summing up the smallest entries of β that there exists some constant $c > 0$ so that

$$2\|\beta(K^c)\|_{\ell_1} \le cs_{q,n}\, r_n^{(q-1)/q},$$

and hence similarly as before

$$\|\gamma\|_{\ell_1} = \|\gamma(K^c)\|_{\ell_1} + \|\gamma(K)\|_{\ell_1} \le 2\|\gamma(K)\|_{\ell_1} + cs_{q,n}r_n^{(q-1)/q} \le 2\sqrt{r_n}\|\gamma\|_{\ell_2} + cs_{q,n}r_n^{(q-1)/q}, \quad (39)$$

where the last inequality follows due to $\gamma(K)$ having at most r_n non-zero entries, essentially by definition of K. Note that the last inequality holds for any sequence r_n. We are going to choose r_n further below in a way that minimzes the bound on the bias.

Just as in the proof of Lemma 1, let $|\gamma_{(1)}| \ge |\gamma_{(2)}| \ge \ldots \ge |\gamma_{(p)}|$ be the ordered entries of γ, let u_n be some sequence for $n \to \infty$ and define $U = \{k : |\gamma_k| \le |\gamma_{(u_n)}|\}$. Analogously to the proof of Lemma 1, we obtain the bound (24) in slightly different notation as

$$\gamma^T C\gamma \ge \phi_{\min}(u_n)\|\gamma\|_{\ell_2}^2 - 2\sqrt{\phi_{\max}(n)\phi_{\min}(u_n)}\|\gamma\|_{\ell_2}\|\gamma(U^c)\|_{\ell_2}. \quad (40)$$

By the same argument as in the proof of Lemma 1, it also holds that $\gamma^T C\gamma \le \lambda\|\gamma\|_{\ell_1}/n$. Just as in (25), we have additionally

$$\|\gamma(U^c)\|_{\ell_2} \le \|\gamma\|_{\ell_1}/\sqrt{u_n},$$

and hence

$$\lambda_n \|\gamma\|_{\ell_1}/n \geq \phi_{\min}(u_n)\|\gamma\|_{\ell_2}^2 - 2\sqrt{\phi_{\max}(n)\phi_{\min}(u_n)}\|\gamma\|_{\ell_2}\|\gamma\|_{\ell_1}/\sqrt{u_n},$$

which is equivalent to

$$\phi_{\min}(u_n)\|\gamma\|_{\ell_2}^2 \leq \|\gamma\|_{\ell_1}(\lambda_n/n + 2\sqrt{u_n^{-1}\phi_{\min}(u_n)\phi_{\max}(n)}\|\gamma\|_{\ell_2}). \tag{41}$$

We have still complete freedom in choosing the sequences $\{u_n\}_{n=1,\ldots,\infty}$ and $\{r_n\}_{n=1,\ldots,n}$. We now choose $u_n = m_{\lambda_n}$, where m_{λ_n} is given in (8) with $m_{\lambda_n} \asymp n^2\lambda_n^{-2}$. By assumption, there exists some $\kappa > 0$ such that $\liminf_{n\to\infty} \phi_{\min}(m_{\lambda_n}) > \kappa$ and the last equation (41) implies thus that there exists a constant $c(\kappa) > 0$ so that for sufficiently large value of n,

$$\|\gamma\|_{\ell_2}^2 \leq c(\kappa)\frac{\lambda_n}{n}\|\gamma\|_{\ell_1}(1 + \|\gamma\|_{\ell_2}).$$

As long as we focus on cases for which $\|\gamma\|_{\ell_2}$ stays bounded (it will turn out below that this holds true), it follows that

$$\|\gamma\|_{\ell_2}^2 = O(\frac{\lambda_n}{n}\|\gamma\|_{\ell_1}).$$

Combing with the bound (39) on the ℓ_1-norm of γ then yields

$$\|\gamma\|_{\ell_2}^2 = O\big(\frac{\lambda_n}{n}\sqrt{r_n}\|\gamma\|_{\ell_2} + \frac{\lambda_n}{n}s_{q,n}r_n^{(q-1)/q}\big).$$

We can still choose r_n freely. We choose r_n to make both terms on the right hand side of the last equation of the same order. Using in particular $r_n \asymp (ns_{q,n}/\lambda_n)^q$, it follows with $m_{\lambda_n} \asymp n^2/\lambda_n^2$ and the definition (12) of the effective sparsity as $s_n^{\text{eff}} = s_{q,n}^{2q/(2-q)}$ that

$$\|\gamma\|_{\ell_2}^2 = O\Big\{\frac{s_{q,n}^q}{m_{\lambda_n}^{1-q/2}}\Big\} = O\Big\{\Big(\frac{s_n^{\text{eff}}}{m_{\lambda_n}}\Big)^{1-q/2}\Big\},$$

which completes the proof. □

5 Numerical Illustration: Frequency Detection

Instead of extensive numerical simulations, we would like to illustrate a few aspects of Lasso-type variable selection if the *irrepresentable condition* is not fulfilled. We are not making claims that the Lasso is superior to other methods for high-dimensional data. We merely want to draw attention to the fact that (a) the Lasso might not be able to select the correct variables but (b) comes nevertheless close to the true vector in an ℓ_2-sense.

An illustrative example is frequency detection. It is of interest in some areas of the physical sciences to accurately detect and resolve frequency components; two examples are variable
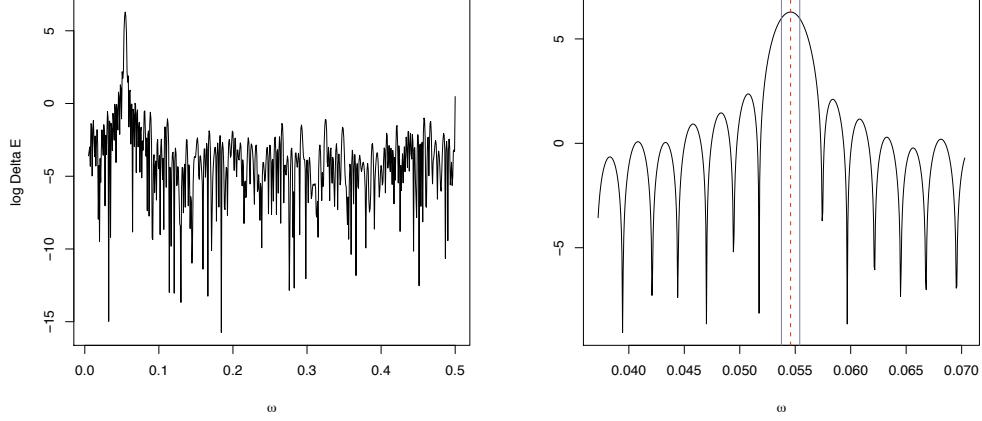
Figure 1: The energy $\log \Delta E(\omega)$ for a noise level $\sigma = 0.2$ is shown on the left for a range of frequencies ω. A close-up of the region around the peak is shown on the right. The two frequencies ω_1 and ω_2 are marked with solid vertical lines, while the *resonance* frequency $(\omega_1 + \omega_2)/2$ is shown with a broken vertical line.

stars (Pojmanski, 2002) and detection of gravitational waves (Cornish and Crowder, 2005; Umstätter et al., 2005). A non-parametric approach is often most suitable for fitting of the involved periodic functions (Hall et al., 2000). However, we assume here for simplicity that the observations $Y = (Y_1, \ldots, Y_n)$ at time points $t = (t_1, \ldots, t_n)$ are of the form

$$Y_i = \sum_{\omega \in \Omega} \beta_\omega \sin(2\pi\omega t_i + \phi_\omega) + \varepsilon_i,$$

where Ω contains the set of fundamental frequencies involved, and ε_i for $i = 1, \ldots, n$ is independently and identically distributed noise with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. To simplify the problem even more, we assume that the phases are known to be zero, $\phi_\omega = 0$ for all $\omega \in \Omega$. Otherwise one might like to employ the Group Lasso (Yuan and Lin, 2006), grouping together the sine and cosine part of identical frequencies.

It is of interest to resolve closely adjacent spectral lines (Hannan and Quinn, 1989) and we will work in this setting in the following. We choose for the experiment $n = 200$ evenly spaced observation times. There are supposed to be two closely adjacent frequencies with $\omega_1 = 0.0545$ and $\omega_2 = 0.0555 = \omega_1 + 1/300$, both entering with $\beta_{\omega_1} = \beta_{\omega_2} = 1$. As we have the information that the phase is zero for all frequencies, the predictor variables are given by all sine-functions with frequencies evenly spaced between $1/200$ and $1/2$, with a spacing of $1/600$ between adjacent frequencies.

In the chosen setting, the *irrepresentable condition* is violated for the frequency $\omega_m = (\omega_1 + \omega_2)/2$. Even in the absence of noise, this *resonance frequency* is included in the Lasso-

estimate for all positive penalty parameters, as can be seen from the results further below. As a consequence of a violated *irrepresentable condition*, the largest peak in the periodgram is in general obtained for the *resonance frequency*. In Figure 1 we show the periodogram (Scargle, 1982) under a moderate noise level $\sigma = 0.2$. The periodogram shows the amount of energy in each frequency, and is defined through the function

$$\Delta E(\omega) = \sum_i Y_i^2 - \sum_i (Y_i - \hat{Y}_i^{(\omega)})^2,$$

where $\hat{Y}^{(\omega)}$ is the least squares fit of the observations Y, using only sine and cosine functions with frequency ω as two predictor variables. There is clearly a peak at frequency ω_m. As can be seen in the close-up around ω_m, it is not immediately obvious from the periodogram that there are *two* frequencies at frequencies ω_1 and ω_2. As said above, the *irrepresentable condition* is violated for the *resonance frequency* and it is of interest to see which frequencies are picked up by the Lasso estimator.

The results are shown in Figures 2 and 3. Figure 3 highlights that the two true frequencies are with high probability picked up by the Lasso. The *resonance frequency* is also selected with high probability, no matter how the penalty is chosen. This result could be expected as the *irrepresentable condition* is violated and the estimator can thus not be *sign consistent*. We expect from the theoretical results in this manuscript that the coefficient of the falsely selected *resonance frequency* is very small if the penalty parameter is chosen correctly. And it can indeed be seen in Figure 2 that the coefficients of the true frequencies are much larger than the coefficient of the *resonance frequency* for an appropriate choice of the penalty parameter.

These results reinforce our conclusion that the Lasso might not be able to pick up the correct sparsity pattern, but delivers nevertheless useful approximations as falsely selected variables are chosen only with a very small coefficient; this behavior is typical and expected from the results of Theorem 1. Falsely selected coefficients can thus be removed in a second step, either by thresholding variables with small coefficients or using other relaxation techniques. In any case, it is reassuring to know that all important variables are included in the Lasso estimate.

6   Concluding Remarks

It has recently been discovered that the Lasso cannot recover the correct sparsity pattern in certain circumstances, even not asymptotically for p fixed and $n \to \infty$. This shed a little doubt on whether the Lasso is a good method for identification of sparse models for both low- and high-dimensional data.

Here we have shown that the Lasso can continue to deliver good approximations to sparse coefficient vectors β in the sense that the ℓ_2-difference $\|\beta - \hat{\beta}^\lambda\|_{\ell_2}$ vanishes for large sample
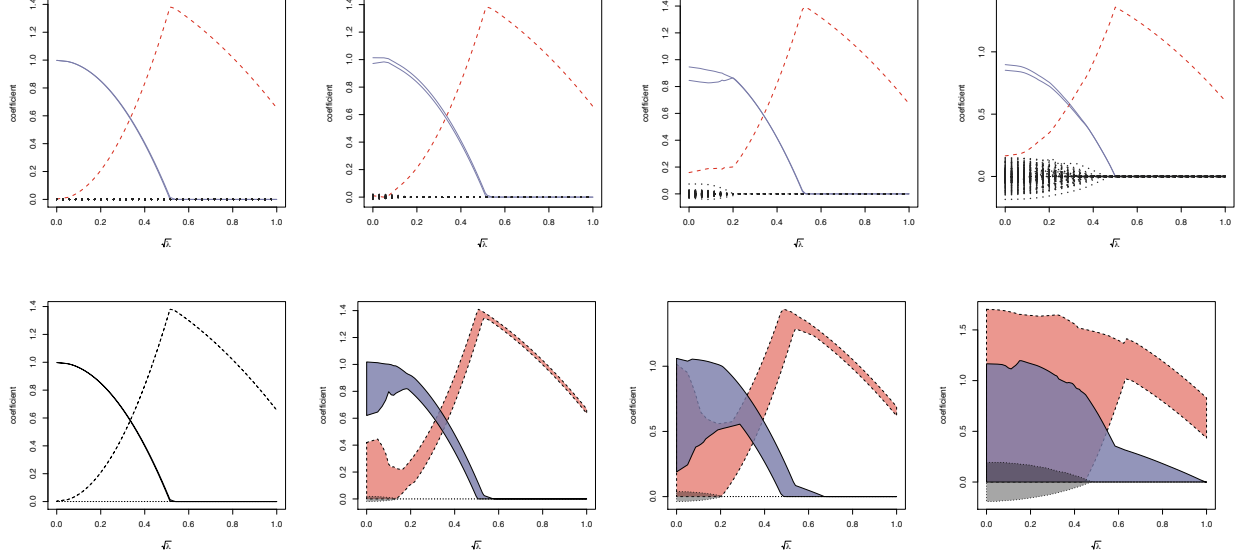
Figure 2: An example where the Lasso is bound to select wrong variables, while being a good approximation to the true vector in the ℓ_2-sense. Top row: The noise level increases from left to right as $\sigma = 0, 0.1, 0.2, 1$. For one run of the simulation, paths of the estimated coefficients are shown as a function of the square root $\sqrt{\lambda}$ of the penalty parameter. The actually present *signal frequencies* ω_1 and ω_2 are shown as solid lines, the *resonance frequency* as a broken line, and all other frequencies are shown as dotted lines. Bottom row: the shaded areas contain, for 90% of all simulations, the regularization paths of the *signal frequencies* (region with solid borders), *resonance frequency* (area with broken borders) and all other frequencies (area with dotted boundaries). The path of the *resonance frequency* displays *reverse shrinkage*, as its coefficient gets in general smaller for smaller values of the penalty. As expected from the theoretical results, if the penalty parameter is chosen correctly, it is possible to separate the *signal* and *resonance* frequencies for sufficiently low noise levels by just retaining large and neglecting small coefficients. It is also apparent that the coefficient of the *resonance frequency* is small for a correct choice of the penalty parameter but very seldom identically zero.
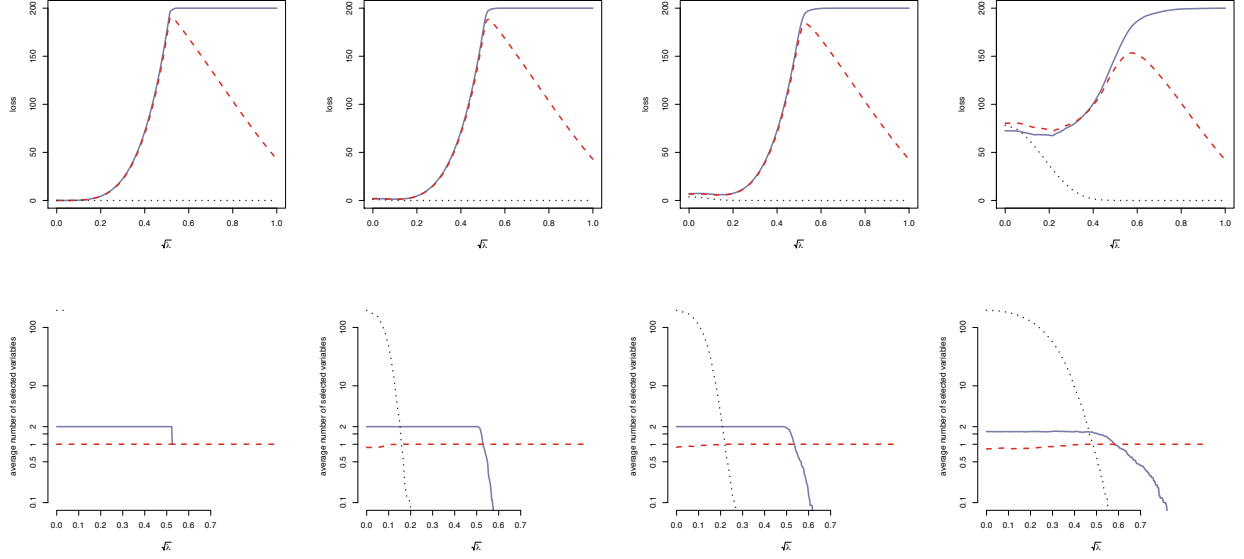
Figure 3: The top row shows the ℓ_2-distance between β and $\hat{\beta}^\lambda$ separately for the *signal* frequencies (solid blue line), *resonance frequency* (broken red line) and all other frequencies (dotted gray line). It is evident that the distance is quite small for all three categories simultaneously if the noise level is sufficiently low (the noise level is again increasing from left to right as $\sigma = 0, 0.1, 0.2, 1$). The bottom row show on the other hand the average number of selected variables (with non-zero estimated regression coefficient) in each of the three categories as a function of the penalty parameter. It is impossible to choose the correct model, as the *resonance* frequency is always selected, no matter how low the noise level and no matter how the penalty parameter is chosen. This illustrates that *sign consistency* does not hold if the *irrepresentable condition* is violated, even though the estimate can be close to the true vector β in the ℓ_2-sense.

sizes n, even if it fails to discover the correct sparsity pattern. The conditions needed for a good approximation the ℓ_2-sense are weaker than the *irrepresentable condition* needed for *sign consistency*. We pointed out that the correct sparsity pattern could be recovered in a two-stage procedure. The first step consists in a regular Lasso fit. Variables with small absolute coefficient are then removed from the model in a second step.

We derived possible scenarios under which ℓ_2-consistency can be achieved as a function of the sparsity of the vector β, the number of samples and the number of variables. The only condition on the design matrix we impose is that singular minimal eigenvalues of the design matrix induced by selecting a small number of variables are bounded away from zero by an arbitrarily small constant.

It was also shown that recovery of sparse vectors β is possible if sparseness is measured in other ways than the number of non-zero entries of β. We obtain recovery of vectors in weak ℓ_q-balls with $0 < q < 1$. In summary, the Lasso is selecting all sufficiently large coefficients, and possibly some other unwanted variables. The number of variables can thus be narrowed down considerably with the Lasso, while keeping all important variables. These results will hopefully support that the Lasso is a useful model identification method for high-dimensional data.

References

Bunea, B., A. Tsybakov, and M. Wegkamp (2006a). Aggregation for gaussian regression. *Annals of Statistics, to appear*.

Bunea, F., A. Tsybakov, and M. Wegkamp (2006b). Sparsity oracle inequalities for the lasso. Technical report.

Candes, E. and T. Tao (2005a). Decoding by linear programming. *Information Theory, IEEE Transactions on 51*(12), 4203–4215.

Candes, E. and T. Tao (2005b). The Dantzig selector: statistical estimation when p is much larger than n. *Arxiv preprint math.ST/0506081*.

Cornish, N. and J. Crowder (2005). LISA data analysis using Markov chain Monte Carlo methods. *Physical Review D 72*(4), 43005.

Davidson, K. and S. Szarek (2001). Local operator theory, random matrices and Banach spaces. In Handbook on the Geometry of Banach spaces, Vol. 1, WB Johnson, J. Lindenstrauss eds.

Donoho, D. (2004). For Most Large Underdetermined Systems of Linear Equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution. Technical report, Department of Statistics, Stanford University.

Donoho, D. (2006). For most large underdetermined systems of linear equations, the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics 59*, 0797–0829.

Donoho, D. and M. Elad (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1-minimization. *PNAS 100*(5), 2197–2202.

Donoho, D., M. Elad, and V. Temlyakov (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on 52*(1), 6–18.

Donoho, D. and X. Huo (2001). Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on 47*(7), 2845–2862.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–451.

Greenshtein, E. and Y. Ritov (2004). Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli 10*(6), 971–988.

Gribonval, R. and M. Nielsen (2003). Sparse representations in unions of bases. *IEEE Transactions on Information Theory 49*(12), 3320–3325.

Hall, P., J. Reimann, and J. Rice (2000). Nonparametric estimation of a periodic function. *Biometrika 87*(3), 545–557.

Hannan, E. and B. Quinn (1989). The resolution of closely adjacent spectral lines. *J. Time Ser. Anal 10*(1).

Joshi, R., V. Crump, and T. Fischer (1995). Image subband coding using arithmetic coded trellis codedquantization. *Circuits and Systems for Video Technology, IEEE Transactions on 5*(6), 515–523.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

LoPresto, S., K. Ramchandran, and M. Orchard (1997). Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework. *Proc. Data Compression Conf*, 221–230.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the waveletrepresentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 11*(7), 674–693.

Meier, L., S. van de Geer, and P. Bühlmann (2006). The group lasso for logistic regression. Technical report, Seminar für Statistik, ETH Zurich.

Meinshausen, N. (2006). Relaxed Lasso. *Computational Statistics and Data Analysis*, to appear.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics 34*, 1436–1462.

Osborne, M., B. Presnell, and B. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*, 319–337.

Paul, D. (2006). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Technical report, UC Davis.

Pojmanski, G. (2002). The All Sky Automated Survey. Catalog of Variable Stars. I. 0 h-6 hQuarter of the Southern Hemisphere. *Acta Astronomica 52*, 397–427.

Scargle, J. (1982). Studies in astronomical time series analysis. II- Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal 263*, 835.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B 58*, 267–288.

Umstätter, R., N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan (2005). LISA source confusion: identification and characterization of signals. *Classical and Quantum Gravity 22*(18), 901.

Valdés-Sosa, P., J. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences 360*(1457), 969–981.

van de Geer, S. (2006). High-dimensional generalized linear models and the lasso. Technical Report 133, ETH Zürich.

Wainwright, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Arxiv preprint math.ST/0605740*.

Yuan, M. and Y. Lin (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, series B, to appear*.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B 68*, 49–67.

Zhang, H. and W. Lu (2006). Adaptive-Lasso for Cox's proportional hazards model. Technical Report 2579, North Carolina State University.

Zhao, P. and B. Yu (2004). Boosted lasso. Technical Report 678, University of California, Berkeley.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. Technical Report 702, Department of Statistics, UC Berkeley, to appear in Journal of Machine Learning Research.

Zou, H. (2005). The adaptive lasso and its oracle properties. Technical Report 645, School of Statistics, University of Minnesota, to appear in Journal of the American Statistical Association.